

03/20/00
JC675 U.S. PTO
09/528889

03-22-00

BOX
PATENT APPLICATION

A
jc675 U.S. PTO
09/528889
03/20/00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Attorney Docket No. 1204

TRANSMITTAL LETTER

Bellevue, Washington 98007

March 20, 2000

TO THE ASSISTANT COMMISSIONER FOR PATENTS:

Transmitted herewith for filing under 137 C.F.R. § 1.53(b) by Express Mail is the complete patent application of: Gregory N. Hullender, John R. Bennett, Patrick M. Haluptzok,

Title: METHOD AND MECHANISM TO REDUCE HANDWRITING RECOGNIZER ERRORS
USING MULTIPLE DECISION TREES

1. This application is a continuation of prior application serial number 08/870,559, filed June 6, 1997, priority from the filing date of which is hereby claimed under 35 U.S.C. § 120.
2. An application consisting of 28 pages of specification and claims and 14 sheets of formal/informal drawings is attached.
3. Please amend the specification Amend the specification by inserting after the title, the following sentence -- This is a continuation of copending application Serial No. 08/870,559, filed on June 6, 1997. -
4. A Preliminary Amendment is enclosed.
5. A copy of the Declaration and Power of Attorney filed in the above noted prior application is attached. [Note: This application is a continuation or divisional, no new matter has been added to this application, and no new inventors have been named.]

 i. Please delete the following inventor(s) named in the prior application, such deleted person(s) not being an inventor of the subject matter claimed herein:

 ii. The entire disclosure of the prior application, from which a copy of the declaration is supplied, is considered as being part of the disclosure of this application and is hereby incorporated by reference herein.
6. The prior application was assigned to Microsoft Corporation, by an assignment recorded at reel 8593/frame 0912.
7. A filing date in accordance with 37 C.F.R. § 1.10 is requested. The Express Mail Certificate appears below.
8. The following additional documents are enclosed: Return Postcard.

COMPUTATION OF FEE

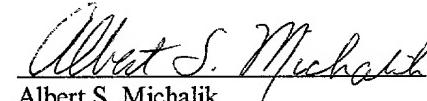
	Number Filed		Number Extra		Rate	Basic Rate
						\$690.00
Total Claims	1 - 20	=	0	X	\$18/\$9	=
Independent Claims	1 - 3	=	0	X	\$78/\$39	=
Multiple Dependent Claims	-0-		---		\$260/\$130	=
TOTAL						\$690.00

- X 9. The Commissioner is hereby authorized to charge any fees under 37 C.F.R. §§ 1.16, 1.17 and 1.18 which may be required during the entire pendency of the application, or credit any overpayment, to Deposit Account No. 500463. This authorization also hereby includes a request for any extensions of time of the appropriate length required upon the filing of any reply during the entire prosecution of this application. A copy of this sheet is enclosed.

Please address all further correspondence to:

MICHALIK & WYLIE, PLLC
 14645 Bel-Red Road
 Suite 103
 Bellevue, WA 98007

Respectfully submitted,
 MICHALIK & WYLIE, PLLC


 Albert S. Michalik
 Registration No. 37,395
 Telephone (206) 653-3520

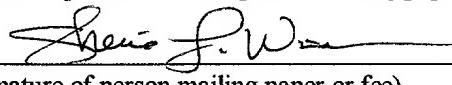
EXPRESS MAIL CERTIFICATE

"Express Mail" mailing label number EK614882412US

Date of Deposit March 20, 2000

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. § 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, Washington, D.C. 20231.

Sherie L. Winans
 (Typed or printed name of person mailing paper or fee)


 (Signature of person mailing paper or fee)

S P E C I F I C A T I O N

TO ALL WHOM IT MAY CONCERN:

Be it known that we, Gregory N. Hullender, a citizen of the United States, residing at 12211 NE 97th Street, Kirkland, Washington 98033, John R. Bennett, a citizen of the United States, residing at 20724 NE 38th Street, Redmond, Washington 98053, and Patrick M. Haluptzok, a citizen of the United States, residing at 15244 NE 73rd Street, Redmond, Washington 98052, have invented a certain new and useful **Method and Mechanism to Reduce Handwriting Recognizer Errors Using Multiple Decision Trees** of which the following is a specification.

**Method and Mechanism to Reduce Handwriting Recognizer Errors
Using Multiple Decision Trees**

FIELD OF THE INVENTION

5 The invention relates generally to the input of user information into computer systems, and more particularly to the recognition of handwritten characters input by a user.

BACKGROUND OF THE INVENTION

10 One of the biggest problems in handwriting recognition technology is reducing the error rate. One frequent type of error results when a user electronically enters a handwritten character, known as a chirograph, that closely matches two or more possible characters in a set to which the computer is 15 trying to match the chirograph, i.e., a set of possible code points. Characters which cause the most errors are typically those which are identical to one another except for a single difference that humans can discern, but contemporary recognizers cannot. For example, certain Japanese symbols are 20 substantially identical to one another but for a single, subtle difference.

25 The problem arises in that most handwriting recognition systems use recognizers based on Neural Nets, Hidden Markov Models (HMM) or a K-Nearest-Neighbor (KNN) approach. These systems perform reasonably well at the task of classifying characters based on their total appearance, but where two characters are identical except for a single difference, they

often fail. While attempts have been made to manually code
recognizers to discern between particularly troublesome pairs,
there are many sets of characters which are easily confused for
one another. This makes the coding process very labor
5 intensive and tedious. Moreover, the result of the coding
depends on one or more person's best guesses as to what to test
for to distinguish the characters. This is not necessarily
very optimal, as there are many possibilities for what best
differentiates two (or more) close characters. Indeed, the
10 best of such systems do not substantially reduce the error
rate. Lastly, each time the recognizer is changed, the set of
characters which are confused by the recognizer also changes,
requiring that much of the labor-intensive coding process be
repeated.

15 Another type of recognition system, based on Decision
trees, especially Classification and Regression Trees (CART),
has been attempted for handwriting recognition. These types of
systems have been rejected because they are unable to make
reliable decisions from among large numbers of characters. By
20 way of example, for systems using Japanese character sets, 6650
different characters are supported. As can be appreciated,
developing a binary tree that can receive any one of 6550
characters and test that character repeatedly and properly down
appropriate branches until a single correct result is found
25 would be an extremely difficult and massive undertaking.

OBJECTS AND SUMMARY OF THE INVENTION

Accordingly, it is a general object of the present invention to provide an improved method and mechanism for
5 reducing the error rate in handwriting recognition.

In accomplishing that object, it is a related object to provide a method and mechanism that differentiates ordinarily-confused characters with a high rate of success.

Another object is to provide a method and system of the
10 above kind that can be automatically trained using sample data.

Yet another object is to provide a method and mechanism of the above kind that is fast, reliable, cost-efficient, flexible and extensible.

Briefly, the present invention provides a method and
15 mechanism for recognizing chirographs input into a computer system. A primary recognizer is provided for converting chirographs to code points, and secondary recognizers (e.g., CART trees) are developed and trained to differentiate chirographs which produce selected code points. Each such
20 secondary recognizer is associated with each selected code point. When a chirograph is received, the chirograph is provided to the primary recognizer whereby a code point corresponding thereto is received. A determination is made as to whether the code point corresponds to one of the selected
25 code points having a secondary recognizer associated therewith.

If not, the code point provided by the primary recognizer is returned. If so, the chirograph is passed to the secondary recognizer, and a code point is returned from the secondary recognizer.

- 5 Other objects and advantages will become apparent from the following detailed description when taken in conjunction with the drawings, in which:

BRIEF DESCRIPTION OF THE DRAWINGS

10 FIGURE 1 is a block diagram representing a computer system into which the present invention may be incorporated;

FIG. 2 is a block diagram representing functional components for training a primary handwriting recognizer according to one aspect of the invention;

15 FIG. 3 is a block diagram representing functional components for sorting chirographs as recognized by a primary recognizer into code point-based files to develop a secondary recognition system according to the present invention;

FIG. 4 represents the contents of an exemplary file sorted
20 by the primary recognizer in FIG. 3;

FIG. 5 is a flow diagram representing the general steps taken to sort the chirographs;

FIG. 6 is a block diagram representing functional components for generating the secondary recognition system from
25 the files of FIG. 3;

FIGS. 7A - 7C comprise a flow diagram representing the general steps take to construct and train the secondary recognition system;

5 FIG. 8 is a block diagram representing functional components for optimizing the recognition mechanism of the present invention;

FIGS. 9A - 9C comprise a flow diagram representing the general steps taken to optimize the recognition mechanism of the present invention;

10 FIG. 10 is a block diagram representing functional components for using the recognition mechanism of the present invention to recognize a chirograph; and

15 FIG. 11 is a flow diagram representing the general steps taken when using the recognition mechanism of the present invention to recognize a chirograph.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Turning to the drawings and referring first to FIG. 1, there is shown a computer system generally designated 20 into which the present invention may be incorporated. The computer system 20 includes a processor 22 operatively connected to storage 24, the storage including random access memory (RAM) 26 and non-volatile storage 28 such as a hard disk-drive, optical drive or the like. As can be appreciated, the non-volatile storage can be used in conjunction with the RAM to provide a

relatively large amount of virtual memory via well-known swapping techniques.

The processor 22 also connects through I/O circuitry 32 to one or more input devices 30, such as a keyboard and pointing device such as a mouse, and a pen-tablet, touch device or other means of getting electronic ink. The system 20 also includes at least one local output device 34 connected to the I/O circuitry 32 for communicating information, such as via a graphical user interface, to the user of the system 20. An operating system is loaded in the storage 24.

In accordance with one aspect of the present invention, and as described in more detail below, those chirographs which often confuse a recognizer are provided to a secondary recognition process. To this end, for each input chirograph, a conventional (primary) recognizer outputs a code point. Instead of directly returning the code point, however, the code point is first examined to determine if it corresponds to a confusion set, i.e., one of two (or more) code points indicative of chirographs which are often confused for each other. If the returned code point does not belong to a confusion set, the code point originally returned by the primary recognizer is returned by the mechanism. However, if the code point indicates a confusion set, a secondary recognizer, specifically developed to distinguish that particular confusion set, is given the chirograph. The

secondary recognizer analyzes the chirograph using more directed tests than performed by the primary recognizer, and returns one of the two (or more) code points based on the results of the tests. Note that such often-confused
5 chirographs are not limited to sets of two, but are often confused with two or more other chirographs.

In an alternative embodiment, the primary recognizer can be trained to recognize shape classes that represent code points (or subsets of codepoints) that look alike. When
10 provided with a chirograph, the primary recognizer thus returns at least one shape class index. The secondary recognizer then determines from the shape class index which code point the chirograph represents. Note that a shape class index is a more general concept, i.e., a code point is a particular type of
15 shape class index. However, for purposes of simplicity, the invention will be described with respect to a primary recognizer that returns code points, except where otherwise noted.

Thus, it should be noted that there are two distinct
20 aspects of the present invention. A first aspect involves the development of the improved recognition mechanism of the present invention using handwriting sample data taken from a number (preferably a large number such as thousands) of users. A second aspect involves the use of a recognition mechanism,
25 developed according to the first aspect of the invention, to

convert a chirograph into a code point. The first aspect, the development of the recognition mechanism, is ordinarily performed in a development environment on a relatively high powered computer system, which may be connected via a network connection or the like to large databases of sample data. The second aspect, the use of the recognition mechanism, is typically performed on a hand-held (palm-top) computing device or the like. Such a device preferably runs under the Windows CE operating system loaded in the storage 24, and includes a touch-sensitive liquid crystal display screen for inputting handwritten characters (chirographs). Other preferred systems include tablet-based desktop personal computers running under the Windows 95 or Windows NT operating systems.

Turning to the development of the recognition mechanism with initial reference to FIGS. 2 - 4, a first training set 40 of sample characters is used by a construction / training process 42 to develop and train a primary recognizer 44 (FIG. 3). A training set is a file including chirographs stored in conjunction with their actual, correct code points, i.e., the code points identifying the character that the user intended to write. The primary recognizer 44 is preferably one which uses a nearest neighbor (KNN) approach. Such recognizers and their construction and training are known and are thus not described in detail hereinafter, except to note that the recognizer 44 matches chirographs to trained recognizer data 46 (such as

endpoint information) to output a code point. Also, it is noted that the recognizer 44 actually may return a (probability-ranked) list of alternative code points in response to a chirograph input thereto, however for purposes of 5 simplicity the present invention will be described with reference to a single returned code point unless otherwise noted. Note that the primary recognizer can also be of the type that returns any type of shape index, with shape codes used to train the primary recognizer. As can be appreciated by 10 those skilled in the art, this technique will work equally well for any recognizer or pattern matching technique that returns discrete proposals.

According to one aspect of the present invention and as best shown beginning in FIG. 3, the primary recognizer 44 is 15 used to begin constructing the secondary recognition mechanisms described above. In general, a sorting process 47 sorts chirographs according to whatever code point (or other shape index) the primary recognizer 44 returns for that chirograph, whereby the way in which the chirographs are sorted with 20 respect to their actual code points ultimately reveals the chirographs that the primary recognizer 44 tends to confuse. To this end, once the primary recognizer 44 has been constructed (FIG. 2), a second training set 48 containing sample chirographs, each stored along with its actual code 25 point, is provided to the primary recognizer 44. Using its

trained recognizer data 46, the primary recognizer 44 returns a code point to the sorting process 47, which sorts the chirographs and actual code points into various files 50₁ - 50_n. Note that if other types of shape indexes are being used, 5 the chirographs are similarly sorted into files for each shape index based on the shape index returned by the primary recognizer 44.

More particularly, as shown in FIG. 5, at step 500 the sorting process 47 first creates a separate file for each code point that is to be supported by the recognition mechanism. 10 The first chirograph in the second training set 48 is selected at step 502, and sent to the primary recognizer 44 at step 504. At step 506, a code point, (which may, in fact, be incorrect) 15 is returned by the primary recognizer 44 to the sorting process 47, and at step 508 written to the file that is associated with the returned code point, along with the actual code point known from the training set 48. For example, if a code point identifying an "A" is returned, the sorting process 47 writes 20 the chirograph that provoked the "A" response along with its actual code point (which may not be an "A") into the "A" file 50₂. As is apparent from FIG. 5, steps 510 - 512 repeat the sorting process 47 with a subsequent chirograph until all 25 chirographs in the training set 48 have been sorted in this manner. Note that the process is the same for shape indexes other than code points.

By way of example, FIG. 4 shows the contents of one such file 50_x, with two confused chirographs therein having different actual code points N_x and N_y. At some time during the sorting process, each distinct chirograph (code points N_x or N_y) had been recognized by the primary recognizer 44 as having N_x as its code point. Accordingly, the sorting process 47 wrote each chirograph and its actual code point into the NX file 50_x. Note that if the primary recognizer 44 made no mistakes, all of the files would contain only (chirograph, 5 actual code point) pairs which matched the code point identifying the file. However, no primary recognizer has ever been found to have such accuracy when provided with suitably 10 large training sets.

Once the second training set 48 has been sorted into the files 50₁ - 50_n, the files 50₁ - 50_n are used by a secondary recognizer development/training process 52 to develop the secondary recognizers. As shown in FIG. 6, Classification and Regression Trees (CART, or CART trees) are used as the 15 secondary recognizers. CART trees are binary decision trees described in the text entitled *Classification and Regression 20 Trees*, Breiman, Friedman, Olshen and Stone, Chapman and Hall, (1984), and herein incorporated by reference in its entirety. At this time in the process, one CART tree in a set 54₁ - 54_n will be developed and trained for each code point (or shape 25 index) supported.

- FIGS. 7A - 7C generally describe how each CART tree is developed. First, a list of questions which are believed to be relevant in distinguishing confusion pairs is assembled. Such questions are frequently based on handwriting strokes, such as,
- 5 "how many total strokes in the chirograph?", "what is the length of the first stroke?" and/or "what is the angle of the third stroke with respect to the first stroke?". At this time there is no standard set of questions, only questions based on the experience and observations of those skilled in the art.
 - 10 Note that the questions may be tailored to the stroke count in the chirograph which is known to the system. As will become apparent, the order of the questions is not important. Moreover, the primary recognizer may have provided some featurization information which the construction process can
 - 15 leverage in addition to its own featurization of the ink.

In general, the CART-building process 52 applies all of the questions to all of the samples (in each of the files 50_1 - 50_n) in order to determine and rank which questions best resolve the primary recognizer's confusion for a given file.

- 20 However, before building the CART, a preliminary test is performed by scanning the sample data at step 700 to determine if all of the actual code points in the given file are the same (and match the file). If so, the data in the sample is pure, whereby secondary recognition will not improve the overall

recognition. Accordingly, the CART-building process 52 terminates for such a sample file.

- In the more typical case wherein a lot of sample data is present in a file, however, at least some of the chirographs 5 will have actual code points that do not directly match the code point (and thus the corresponding file) determined by the primary recognizer. Thus, to build a CART, at step 701 of FIG. 7A, the first question in the list is obtained, the first sample chirograph (from one of the files, e.g., 50₁) is 10 obtained at step 702, and at step 704 the question is applied to the sample, producing a result. For example, the question may inquire as to the horizontal length of the first stroke, and result in a value of nine (highest x-coordinate minus lowest x-coordinate equals 9) for the first sample.
- 15 The resulting value is saved in conjunction with the actual code point for that sample, e.g., (value, actual code point) at step 706, and at steps 708 - 710, the process repeated on the next sample in the selected file 50₁. This time through, step 706 again saves whatever (value, actual code 20 point) results from this next sample for this first question. Ultimately, via steps 708 - 710, all the samples in the file will be tested against the first question, whereby all of the resultant values and their associated code points are known in the given file for which the CART tree is being developed for 25 this question.

- At this time, the steps of FIG. 7B are executed, in general to find out which of the values divides (splits) the chirographs in the file along the lines of their associated actual code points. It should be noted that it is possible,
- 5 although generally impractical, to test every conceivable value with each question in a brute-force approach to determine the best split. For example, every length from 1 to 1000 may be tested for the length question, and so on with other wide ranges of values for the other questions. Instead, however,
- 10 only the actual results obtained by the steps of FIG. 7A are used for this purpose, substantially speeding up the split-testing process of FIG. 7B. Moreover, while each unique result can be applied as a binary question against each of the samples to determine the split, a more optimal way is to use the
- 15 already existing result data to determine the best split.

More particularly, step 720 sorts the results obtained for the given question (in FIG. 7A) into an ordered range of (values, actual code points). Thus, for the length example, the shortest lengths may have been forty and the longest one-hundred. To test the quality of the split, at step 722, each of the code points having a "value equal forty" are moved into one (e.g., left) subset, and all code points having other values placed (or left to remain) in another (e.g., right subset). The movement of the code points occurs incrementally.

25 Then, when all of the "forty" values have been moved, at step

- 724, the quality of the split is measured as described below. As can be appreciated, CART trees are binary, and thus the movement effectively asked the question of all samples in the file 50₁, "is the horizontal length of the first stroke less than or equal to forty?", i.e., $X_{\max s1} - X_{\min s1} \leq 40$? Each chirograph is effectively sorted according to this question, with the code points for yes answers placed in one (left branch) set and the code points for no answers placed in another (right branch) set.
- In keeping with the invention, at step 724, the quality of the split is evaluated according to some split criterion. A preferred way to determine the quality of the split is to test for homogeneity of the sets using the Gini diversity index. In general, the Gini diversity index uses a sum of the squares method for the homogeneity (h) using the quantities of the code points in each of the left and right sets, i.e.,
- $$h_{\text{Left}} = [cp_{1\text{Left}} / (cp_{1\text{Left}} + cp_{2\text{Left}} + \dots + cp_{n\text{Left}})]^2 + [cp_{2\text{Left}} / (cp_{1\text{Left}} + cp_{2\text{Left}} + \dots + cp_{n\text{Left}})]^2 + \dots + [cp_{n\text{Left}} / (cp_{1\text{Left}} + cp_{2\text{Left}} + \dots + cp_{n\text{Left}})]^2$$
- and
- $$h_{\text{Right}} = [cp_{1\text{Right}} / (cp_{1\text{Right}} + cp_{2\text{Right}} + \dots + cp_{n\text{Right}})]^2 + [cp_{2\text{Right}} / (cp_{1\text{Right}} + cp_{2\text{Right}} + \dots + cp_{n\text{Right}})]^2 + \dots + [cp_{n\text{Right}} / (cp_{1\text{Right}} + cp_{2\text{Right}} + \dots + cp_{n\text{Right}})]^2,$$
- where subscript left and right indicate the left and right sets and $cp_1, cp_2 \dots cp_n$ represent the number of code points of each code point value in each set. The homogeneity results are then weighted by the number of code points in each set, such that

the final homogeneity equation for this question (Q1) and value (V1) reads:

$$H(Q1, V1) = [(h_{Left})(cp_{1Left} + cp_{2Left}) + (h_{Right})(cp_{1Right} + cp_{2Right})] / (cp_{1Left} + cp_{2Left} + cp_{1Right} + cp_{2Right}).$$

- 5 Step 726 tests the quality of the split against any previous results, if any, and if better, step 727 saves the homogeneity result $H(Q1, V1)$ as the best quality split. Note that step 727 saves the best split over all the questions so far, including possibly the present question, whereby step 726 compares each
10 subsequent split agains the result from the best (question, value) previously determined.

Steps 728 - 730 cause the split for the next value to be tested and compared again, this time using the next value in the range, e.g., forty-one (41). By incrementally moving the
15 code points corresponding to this value to the left split subset at step 722, the sample is now effectively split with code points in the left subset being those having values less than or equal to forty-one. Note that the code-points associated with forty previously moved to the left subset
20 remain there, since these are also less than forty-one. Then, at step 724, the next homogeneity $H(Q1, V2)$ is computed, compared at step 726 (against the value for forty, which was the best so far), and if an improvement, saved as the best value, along with the identity of its corresponding question,
25 at step 727.

Once all the values have been tested in this manner, as determined by step 728, the best value (i.e., the value providing the most homogenous split) for this first question has been determined. Subsequently, the next question is selected (steps 732 - 734), and the process repeated on the samples in the file using this next question. The best (question, value) pair will continue to be saved for comparison against splits of other questions and values, and so on, until the overall best single (question, value) pair is known.

Ultimately, all of the values for all of the questions will have been used to split the sample and calculate the homogeneity therefor, with the best (question, value) pair having been saved. At this time, the steps of FIG. 7C are executed to establish a CART tree which appropriately asks the best questions in the best order according to which produces the best (most homogeneous) split of the code points. Significantly, the process thus automatically locates the best (available) question/value pair at various levels. As a result, the process often finds that the best available question is one that is not originally believed by humans to be that important, but was nevertheless submitted. Note that this is not necessarily the ideal question and value combination for homogeneously splitting the sample set, only the best question and value of those available.

In any event, once the best question and value is known, the sample set (file 50₁) is then split at step 742 into two subsets using this best question/value pair. Then, as represented by step 744, the process is iteratively repeated on 5 each of these two subsets to find the next best question and value pair for most homogeneously splitting each of the subsets. The process is repeated recursively, (i.e., the process returns to step 700 of FIG. 7A to optimally split each of the two subsets), branching into more and more homogeneous 10 subsets until a point is reached at which the homogeneity is no longer improved. As can be appreciated, the recursive operation at lower and lower levels establishes the best question/value pairs at each branch and level to further refine the distinction the confusion pairs. At step 746, a CART tree 15 is built from these question and value pairs at the various levels.

However, primarily because available sample data is limited, the CART trees tend to be imperfect, especially at the lower levels. Moreover, the CART trees may be large, requiring 20 a lot of storage that is not generally available in hand-held computing devices. Accordingly, at step 748 a new set of samples is applied to the CART to test which of its embedded questions are making the correct decisions. Those questions which are determined to be ineffective at resolving the 25 confusion pairs are removed (pruned) from the tree at step 750.

This leaves a more manageable CART in terms of size while not adversely affecting the recognition accuracy.

After all the files $50_1 - 50_n$ have had a CART tree developed therefor in the manner described above, a CART tree exists for each code point supported by the system. However, some of these CART trees may have only one possible output that is the same code point as output by the primary recognizer. Such CART trees do not improve the recognition, and are thus discarded.

Similarly, (if the primary recognizer is one that returns a code point), there is no assurance that a particular CART tree will improve the recognition accuracy of the primary recognizer. This is particularly true when CART trees are developed with lesser amounts of data, as in general, the more training data, the better the CART tree at resolving confusion. If a particular CART tree does not improve the recognition some threshold amount (which may be even a very slight improvement), there is no reason to keep it, since a CART tree costs storage space. Similarly, even though CART trees are extremely fast, secondary recognition using a CART tree adds to the total recognition time, again adding cost.

Thus, FIGS. 8 and 9A - 9C represent one process for optimizing the recognition mechanism by discarding unneeded CART trees. First, at step 900, a first chirograph from a third training set 56 (FIG. 8) is selected. At step 902 the

chirograph is sent to the primary recognizer 44. At step 904, if the code point returned by the primary recognizer 44 is the same as the actual, correct code point, a primary recognizer match count 62 for the CART tree (i.e., this file) is
5 incremented at step 906.

Then, at step 908, the appropriate CART tree, corresponding to the code point returned from the primary recognizer 44, is selected. At step 920, the same chirograph is now provided to the CART tree, whereby a decision is made by
10 the CART tree and a code point returned therefor. At step 922, if the code point returned by the CART tree 52 is the same as the actual, correct code point, a CART match count 66 for this CART tree is incremented at step 924. Steps 926 - 928 repeat the process until all chirographs in the third training set 56
15 are tested.

Lastly, FIG. 9C compares the primary and CART match counts for each CART tree to determine if the CART tree improved the recognition. More particularly, the first supported code point (there is one file per each) is chosen at step 940, and the
20 CART match count 66 for this CART tree compared against the primary recognizer match count 62. If the CART match count is less than or equal to the primary match count, the CART tree is discarded at step 944 since it did not improve the recognition mechanism. Otherwise the CART tree for this code point is
25 kept. Steps 946 - 948 repeat the comparison until all

supported code points have been tested. Note that if desired, step 942 can be a more complex test so as to discard any CART tree that does not improve the recognition process by some threshold amount. For example, if the ratio of the CART match 5 count to the primary match count does not exceed 1 plus some fractional amount, then the CART tree can be discarded. However, it should be noted that with the above optimizations, the CART trees (which may number several hundred) only add about 18 kilobytes to a one megabyte primary recognizer, so any 10 memory savings resulting from discarding a CART tree that only rarely improves recognition is probably not worth a reduction in recognition accuracy.

Indeed, the combined primary and secondary recognition mechanism of the present invention has been thoroughly tested, 15 and for certain confusion pairs has a 99.7 percent accuracy rate. The 0.3 percent error rate is believed to result from characters too poorly written even for humans to discern, and in fact is comparable to the recognition error rate of humans. Note that the present invention is highly flexible and 20 extensible. For example, if the primary recognizer changes, thereby changing the confusion pairs, new CART trees can be automatically generated and tested in the above-described manner. Similarly, as additional questions are conceived, those questions can be simply tried against the others to

determine if they improve the homogeneity of a given split and thus the accuracy of the mechanism.

Note that with a primary recognizer that returns a shape index that is something other than a code point, at least some minimal CART tree that maps the shape index to a code point is needed. However, although this increases the amount of CART storage, using shape codes to train the primary recognizer and partition data for generating CART trees, and then using actual code points to train the CART trees, results in a limited, but smaller and faster primary recognizer. Advantages are again obtained because the primary recognizer is optimized to break chirograph data into classes of shapes, while the CART trees are optimized to focus in on fine differences in the chirographs.

Once the recognition mechanism including the CART trees is developed, the recognition mechanism may be used in a relatively low powered system, e.g., a hand-held personal computing device. FIGS. 10 - 11 shown how the system is used to recognize a character. First, the system receives a chirograph 80 (FIG. 10) from a user at step 1100 in a known manner, such as via pen input on a touch-sensitive screen. Then, the recognition mechanism of the present invention submits the chirograph to the primary recognizer 44 and receives a code point (or shape index) 82 therefrom (step 1102). At step 1104, the code point 82 is used (by a lookup

process 84 or the like) to determine if the code point has a CART tree associated therewith. If not, the primary recognizer's returned code point 82 is returned by the recognition mechanism at step 1108 as the returned code point

5 88.

However, if a CART tree is associated with the code point 82, the appropriate CART tree in the set of available CART trees 72 is selected and the chirograph 80 submitted thereto at step 1106. Note that a shape index code that is not by itself

10 a code point has a secondary recognizer (CART tree) associated therewith, even if only a minimal one that converts the shape index to a code point. The code point returned by the selected CART tree is returned at step 1108 as the returned code point 88. The recognition mechanism repeats until the user is done

15 writing, as detected by step 1110.

Although not necessary to the invention, if the primary recognizer returns a list of alternative code points, the list can be scanned for code points having associated CART trees, and the secondary recognizer operated for one or more of the

20 code points in the list. The secondary process reorders the list with the result from the CART tree placed on top, i.e., with the highest probability. Moreover, CART trees can provide alternatives ranked by probabilities, all of which can be weaved into a composite, probability-ranked list.

Moreover, a plurality of CART trees can be associated with a single character. For example, a first CART tree can be provided as a secondary process for differentiating two-stroke "A"-shaped characters, and a second, distinct CART tree for 5 differentiating three-or-more-stroke "A"-shaped characters.

Moreover, the primary recognizer can be arranged to split strokes, e.g., a one-stroke "A" shaped character can first be split into two strokes by the primary recognizer prior to its analysis thereof. Features other than stroke count may 10 similarly be used by the primary and/or secondary recognizers.

Lastly, it can be appreciated that the invention is applicable to both on-line and off-line character recognition. In on-line character recognition, the points in the character are received as coordinates of the form (x, y, time), i.e., the 15 points in sequence along with pen-up and pen-down positions are known. Off-line chirographs are only x-y points in no particular order. The invention is valuable in either type of recognition, although the primary and secondary recognizer (e.g., questions therefor) will be rather different.

20 As can be seen from the foregoing detailed description, there is provided an improved method and mechanism for reducing the error rate in handwriting recognition. The method and mechanism differentiates ordinarily-confused characters with a high rate of success, and can be automatically trained using

sample data. The method and mechanism that is fast, reliable, cost-efficient, flexible and extensible.

While the invention is susceptible to various modifications and alternative constructions, certain

- 5 illustrated embodiments thereof are shown in the drawings and have been described above in detail. It should be understood, however, that there is no intention to limit the invention to the specific forms disclosed, but on the contrary, the intention is to cover all modifications, alternative
- 10 constructions, and equivalents falling within the spirit and scope of the invention.

CONFIDENTIAL - SECURITY INFORMATION

WHAT IS CLAIMED IS:

1. A method of recognizing chirographs input into a computer system, comprising the steps of:

 providing a primary recognizer for converting chirographs to shape indexes;

 providing a plurality of secondary recognizers to convert chirographs into code points, and associating the secondary recognizers with at least some of the shape indexes, each secondary recognizer capable of overriding a shape index provided by the primary recognizer;

 receiving a chirograph;

 providing the chirograph to the primary recognizer and receiving a shape index therefrom;

 determining whether one of the secondary recognizers is associated with the shape index, and if so, selecting that secondary recognizer as a selected secondary recognizer; and

 passing the chirograph to the selected secondary recognizer and returning a code point from the secondary recognizer.

ABSTRACT

An improved method and mechanism for recognizing chirographs (handwritten characters) input into a computer system. A primary recognizer is provided for converting chirographs to code points, and secondary recognizers such as binary CART trees are developed and trained to differentiate chirographs which produce certain code points at the primary recognizer. Each such secondary recognizer is associated with each selected code point. When a chirograph is received, the chirograph is provided to the primary recognizer whereby a code point corresponding thereto is received. If the code point corresponds to one of the secondary recognizers, the chirograph is passed to the secondary recognizer, and a code point is returned from the secondary recognizer. If not, the code point provided by the primary recognizer is returned. The invention sets forth an automated process for training the CART trees and for optimizing the recognition mechanism by discarding CART trees which do not improve on the recognition accuracy of the primary recognizer.

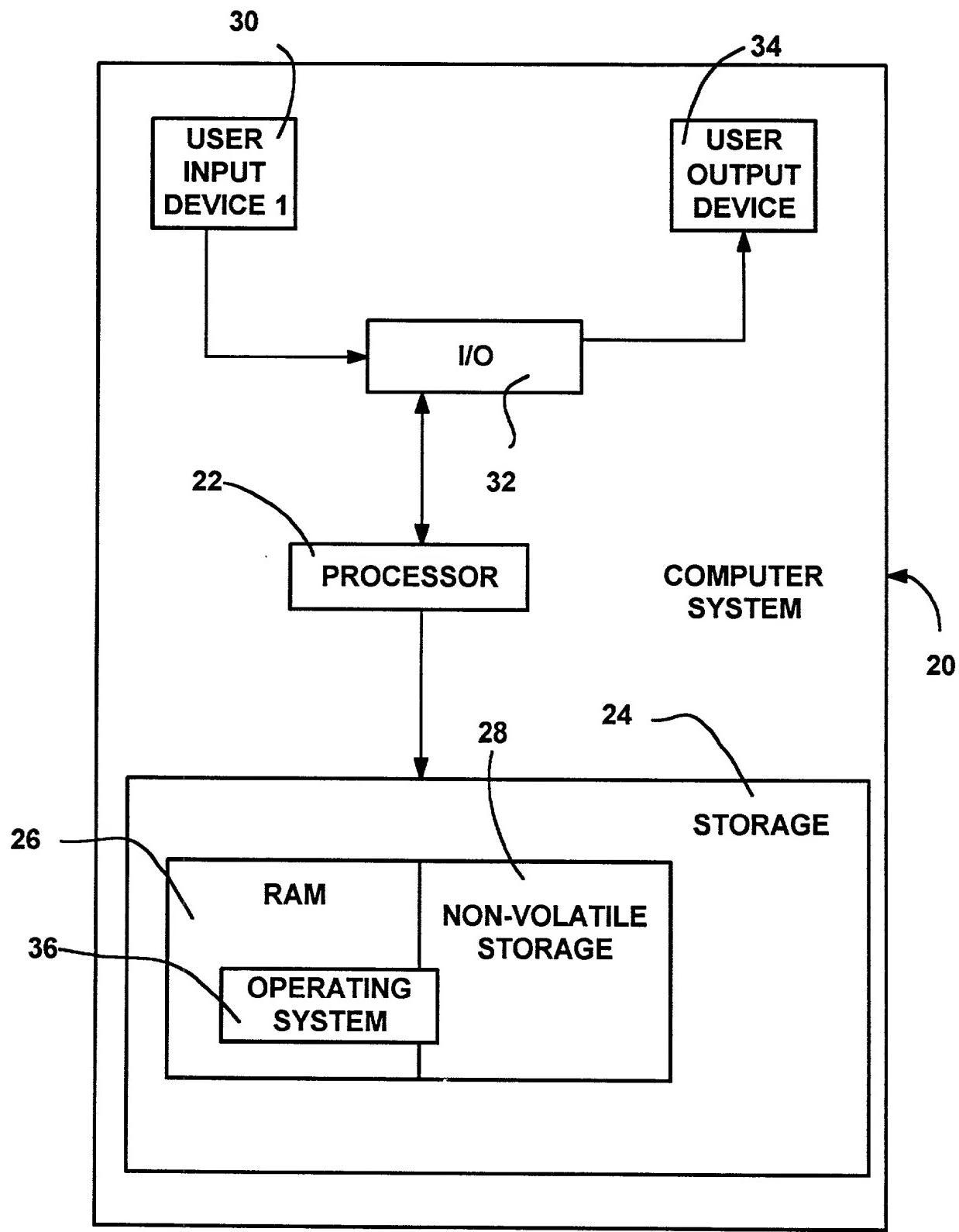


FIG. 1

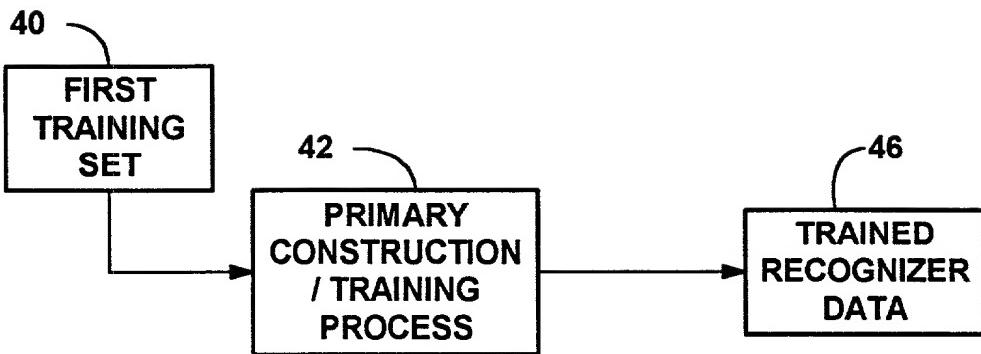


FIG. 2

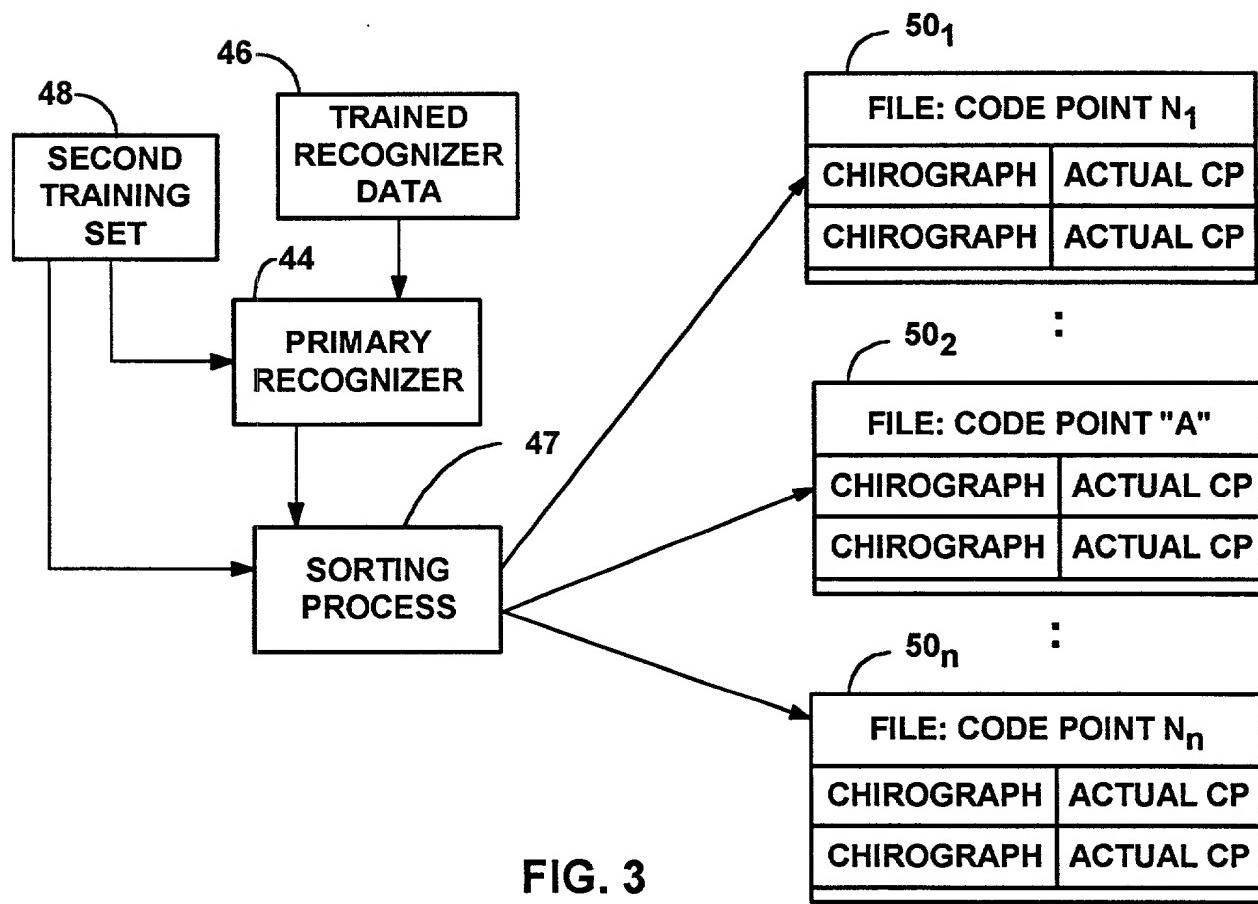


FIG. 3

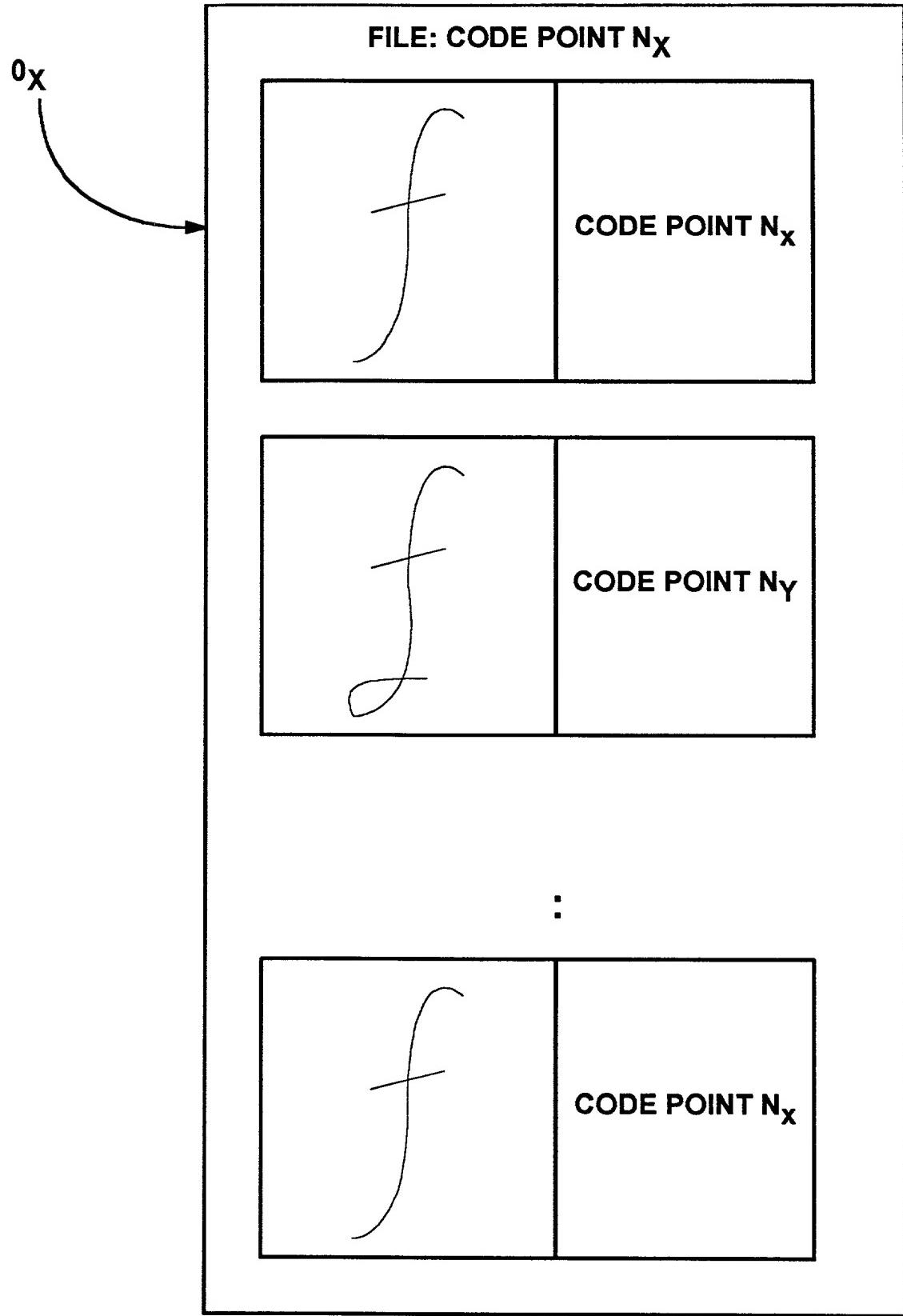
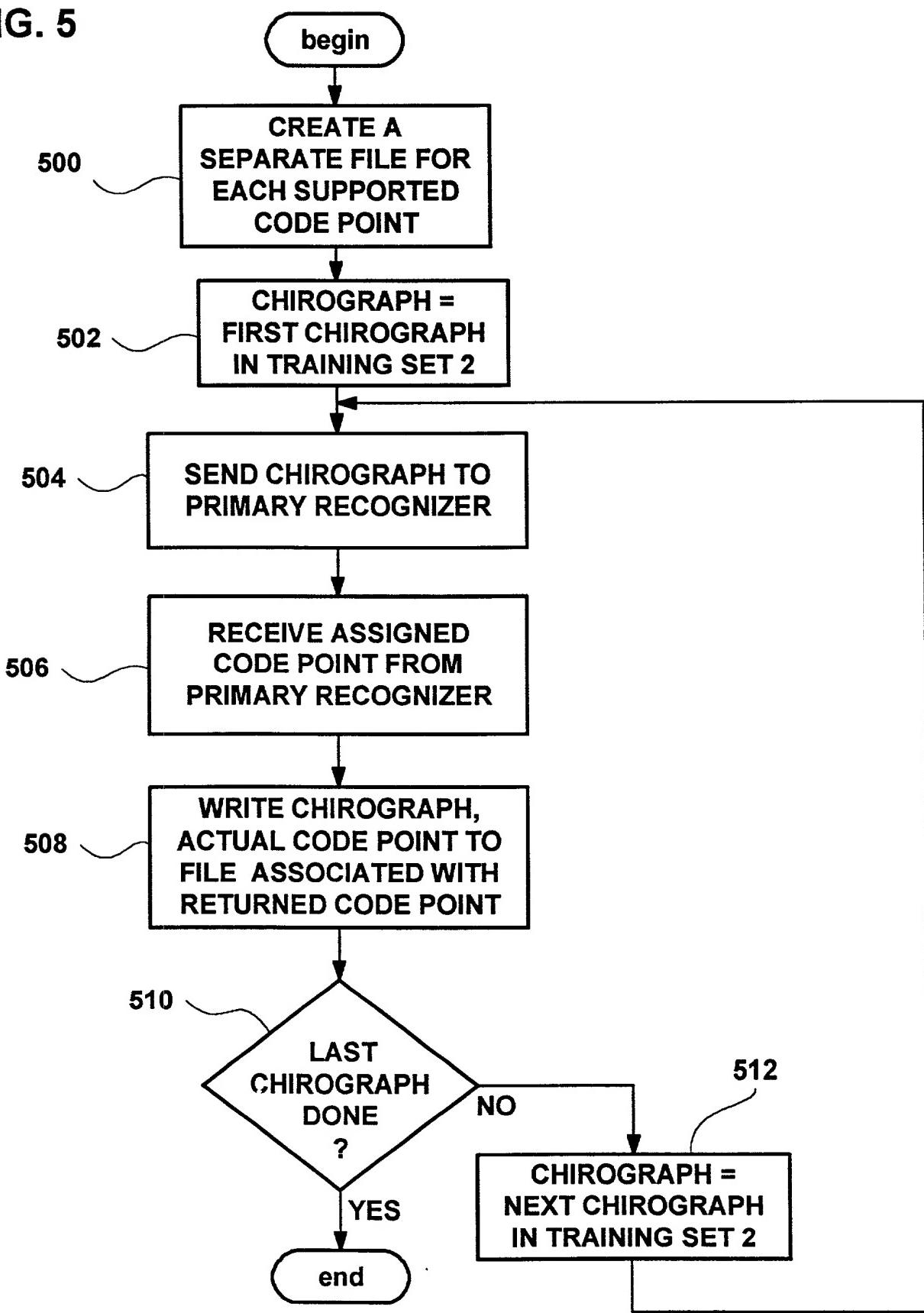


FIG. 4

FIG. 5



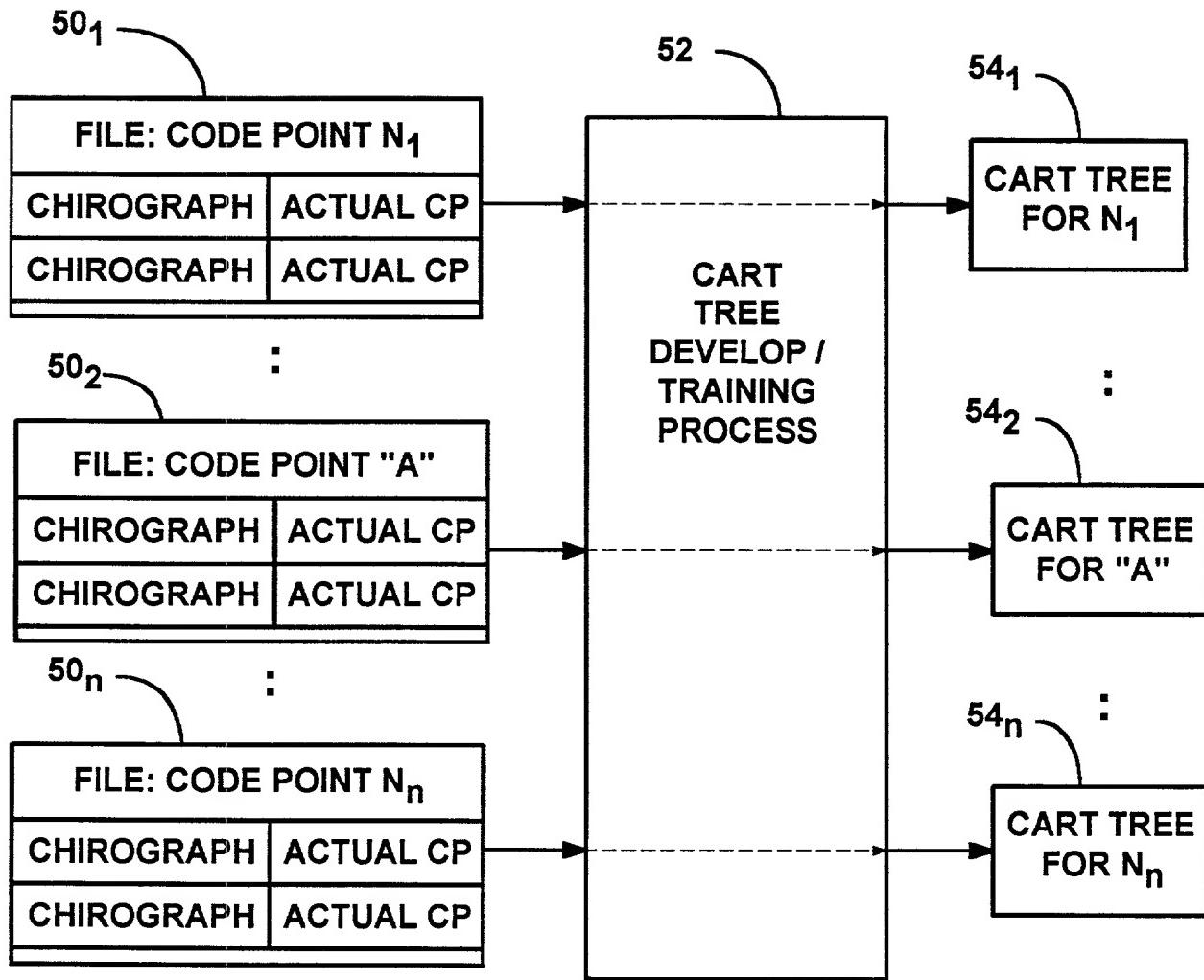


FIG. 6

FIG. 7A

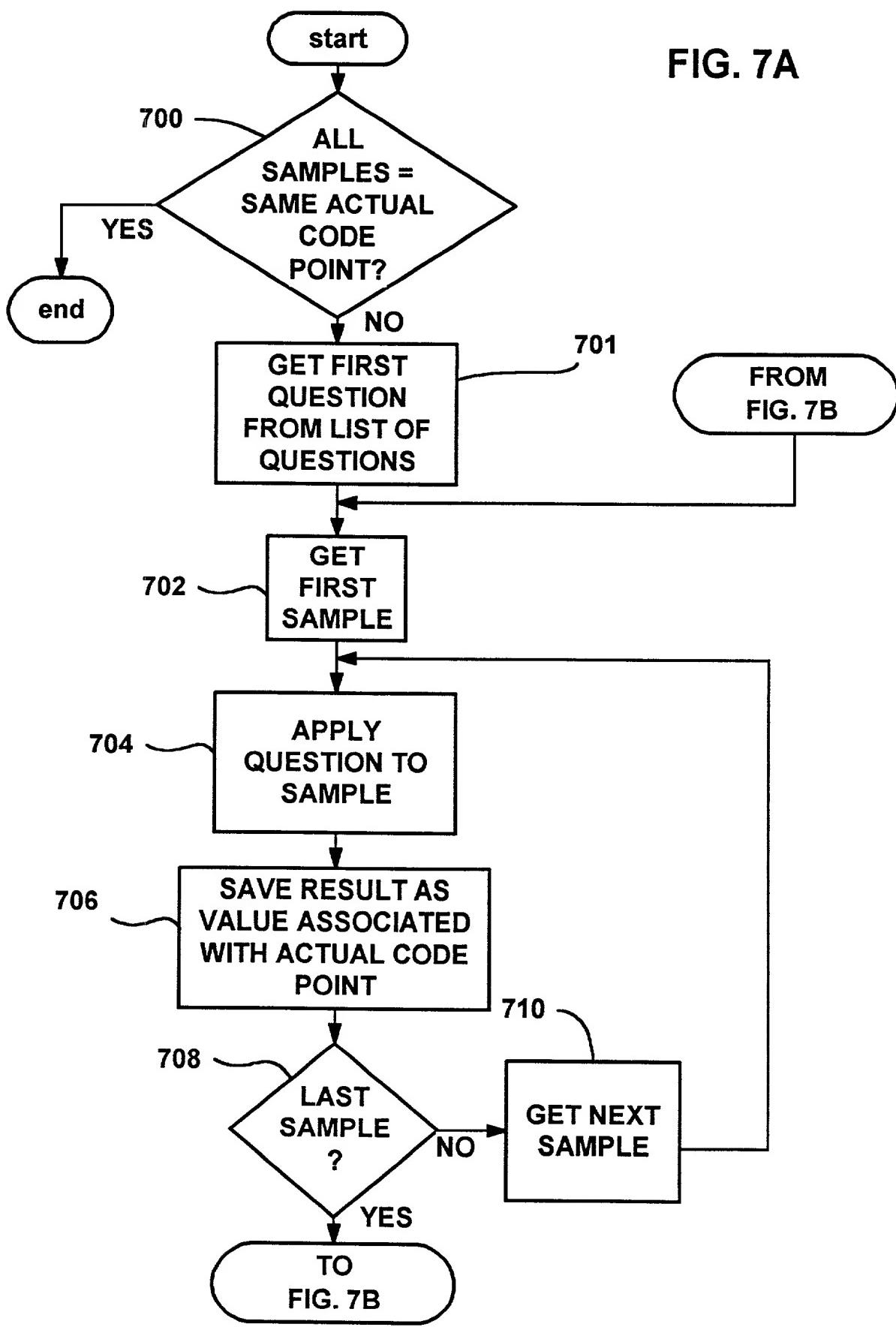


FIG. 7B

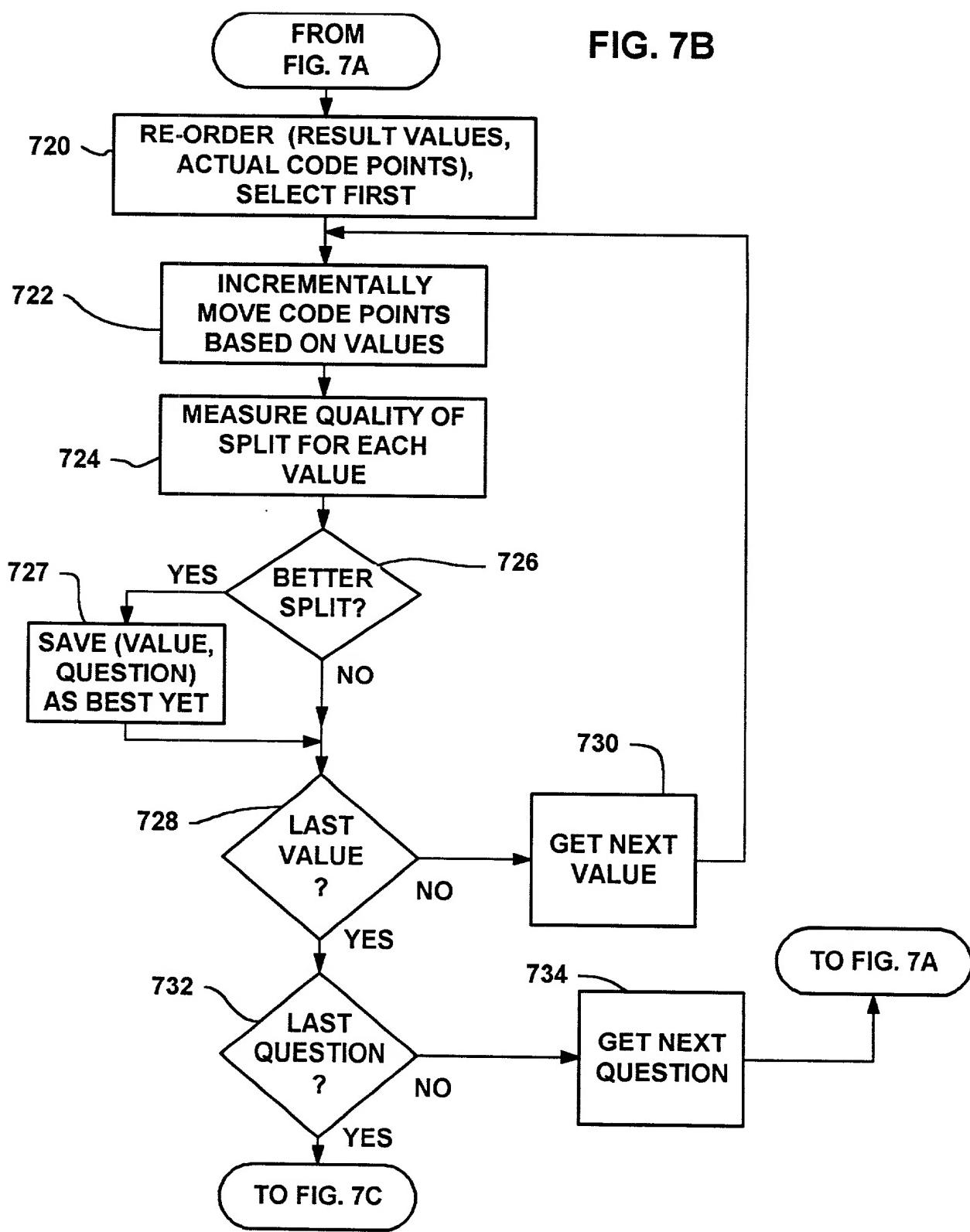


FIG. 7C

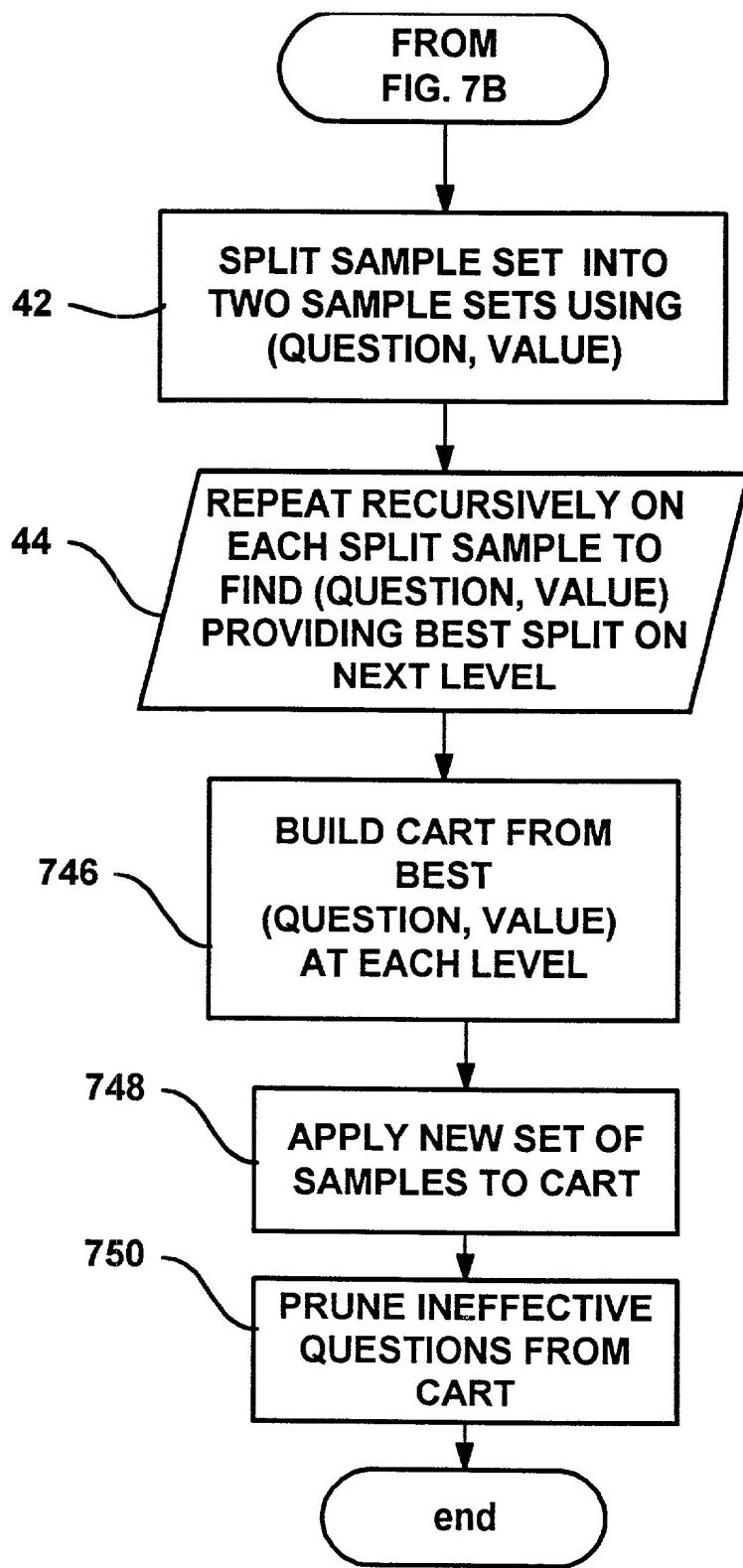


FIG. 8

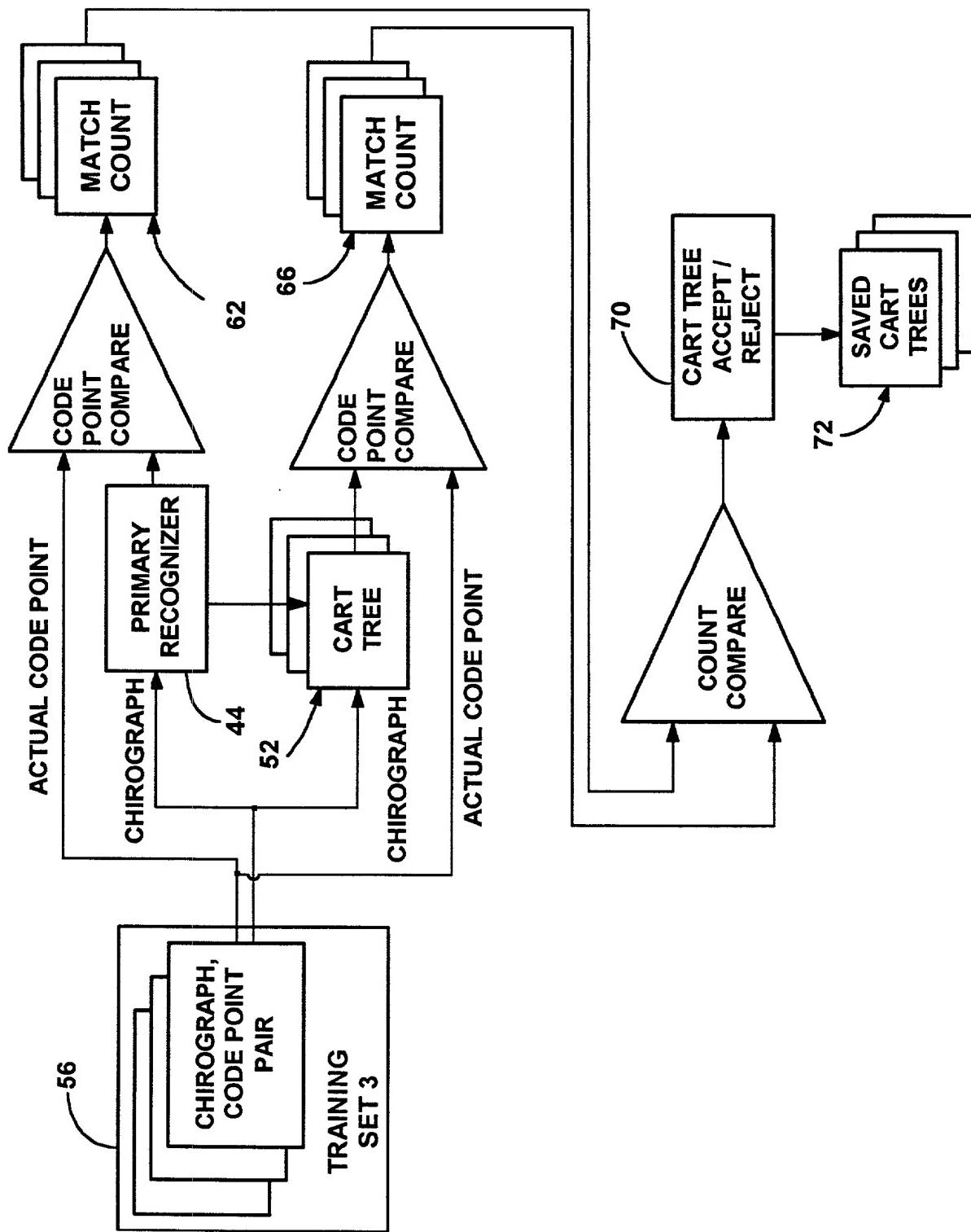


FIG. 9A

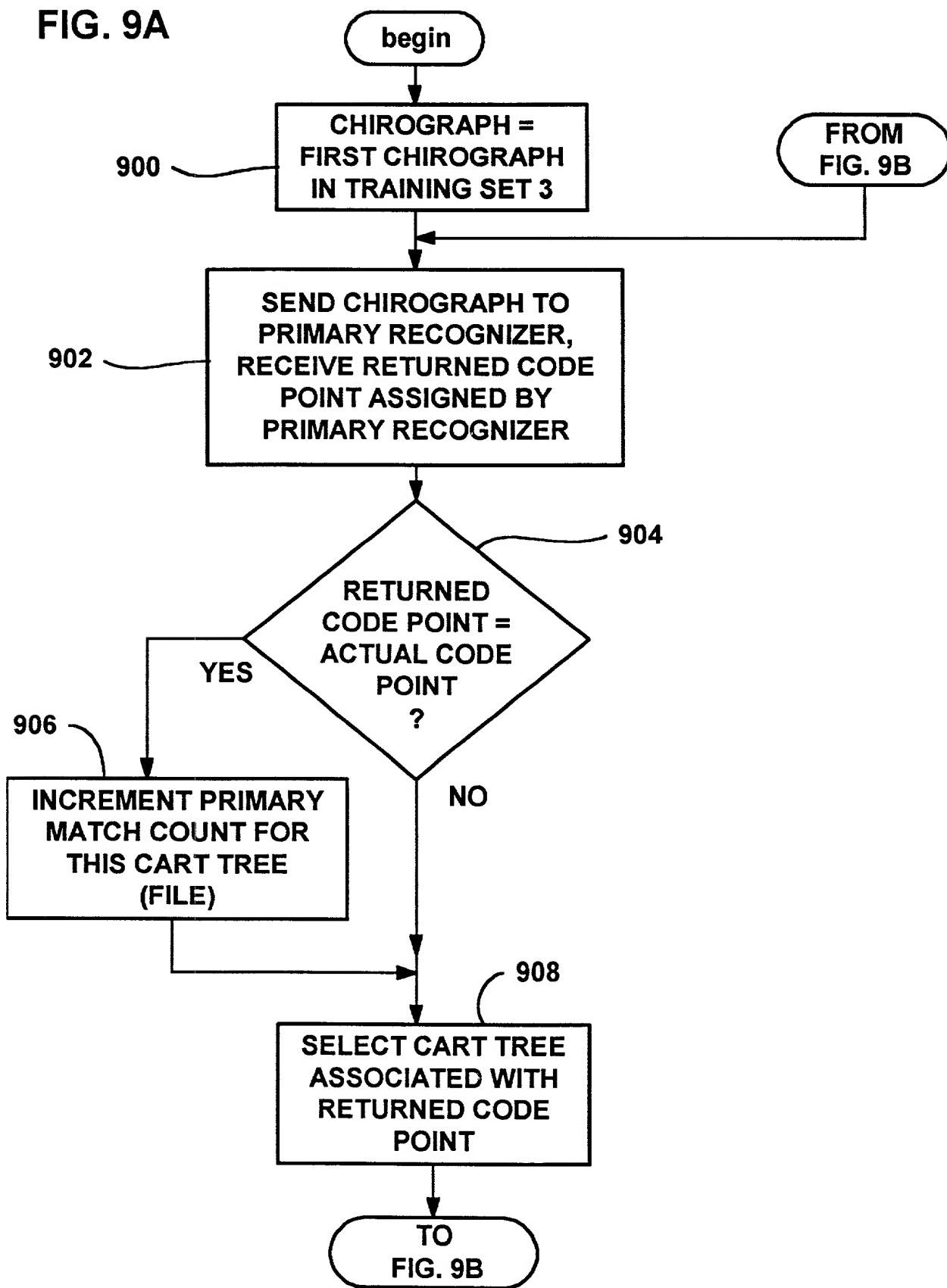


FIG. 9B

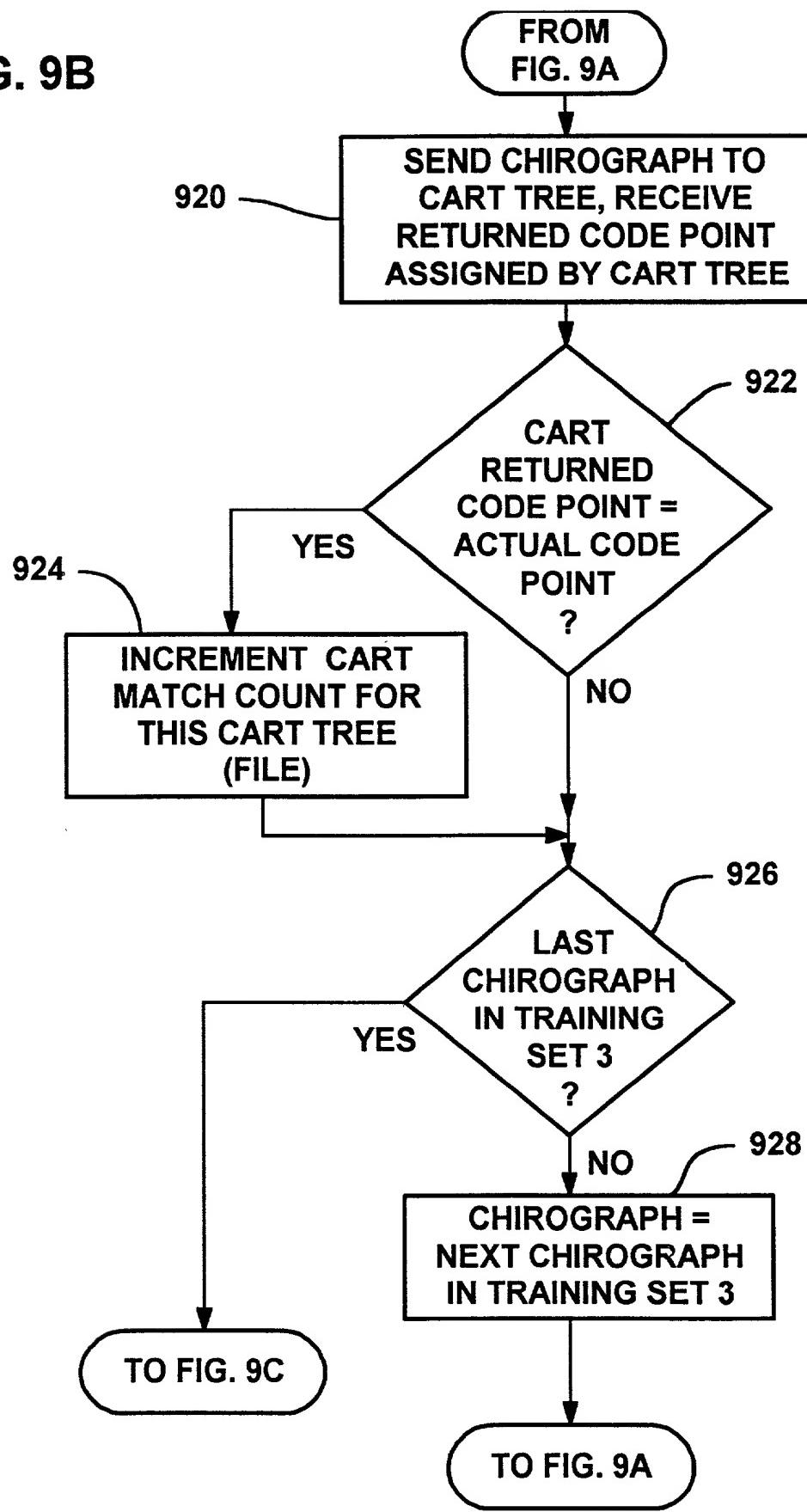
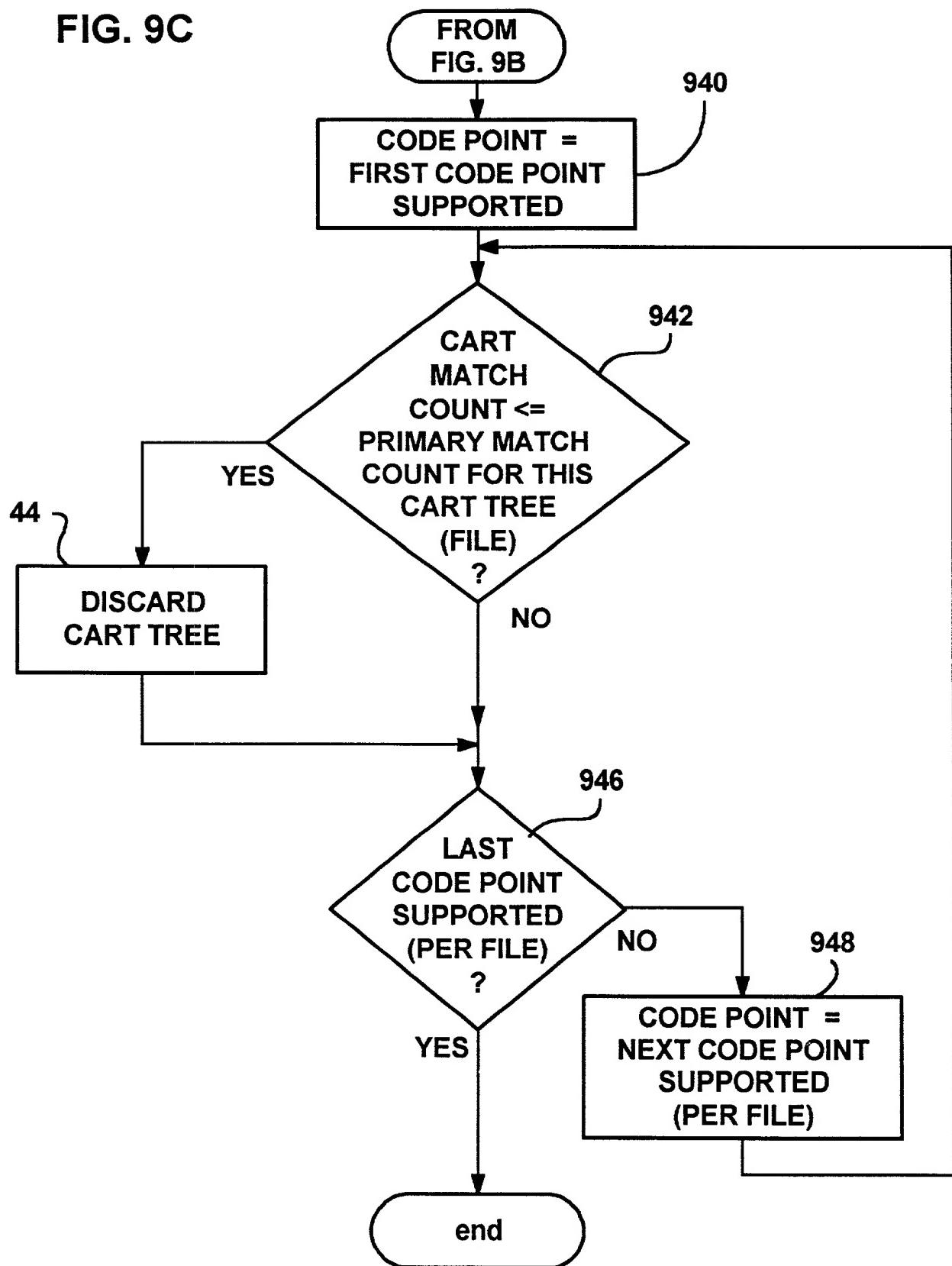


FIG. 9C



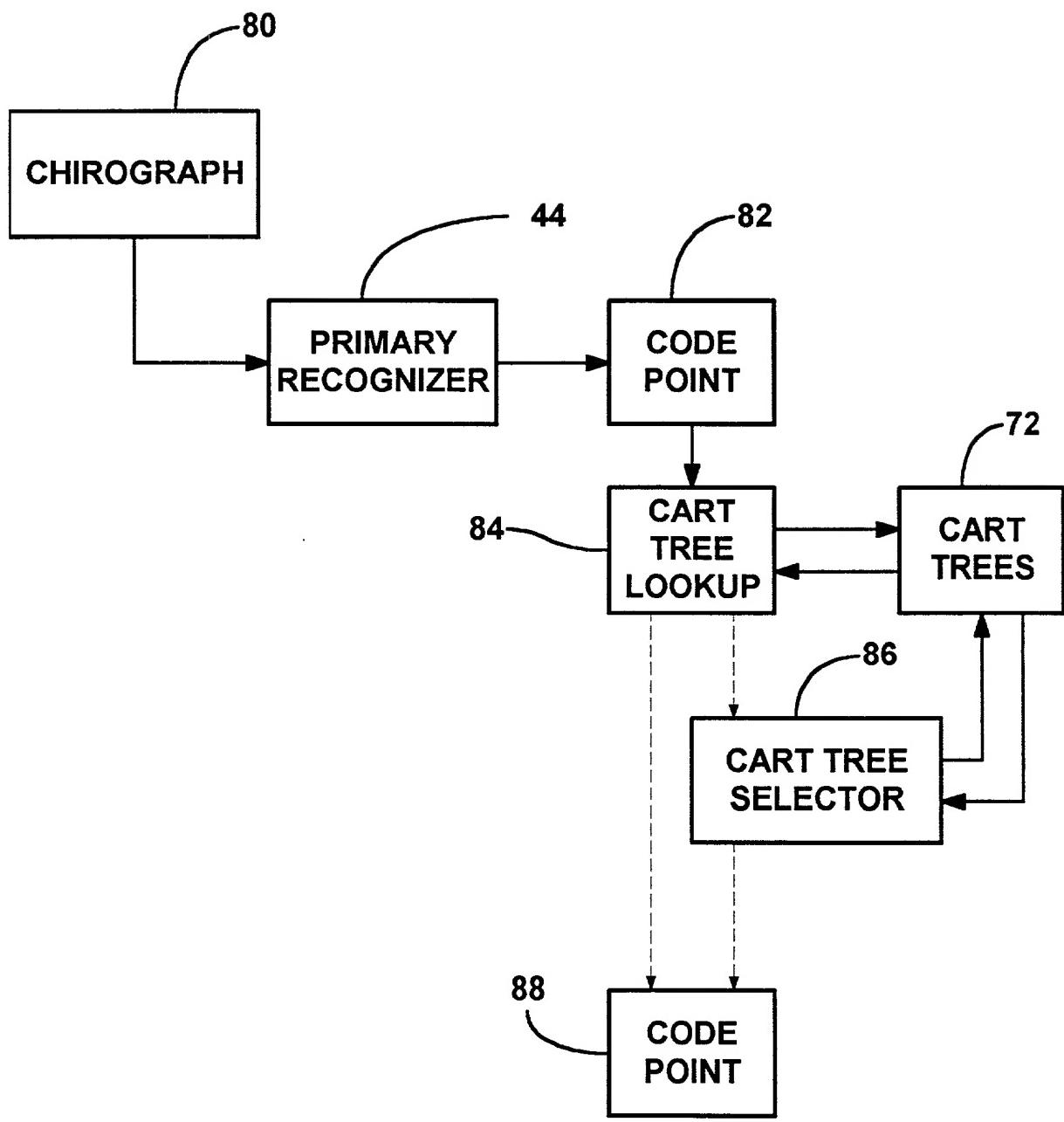
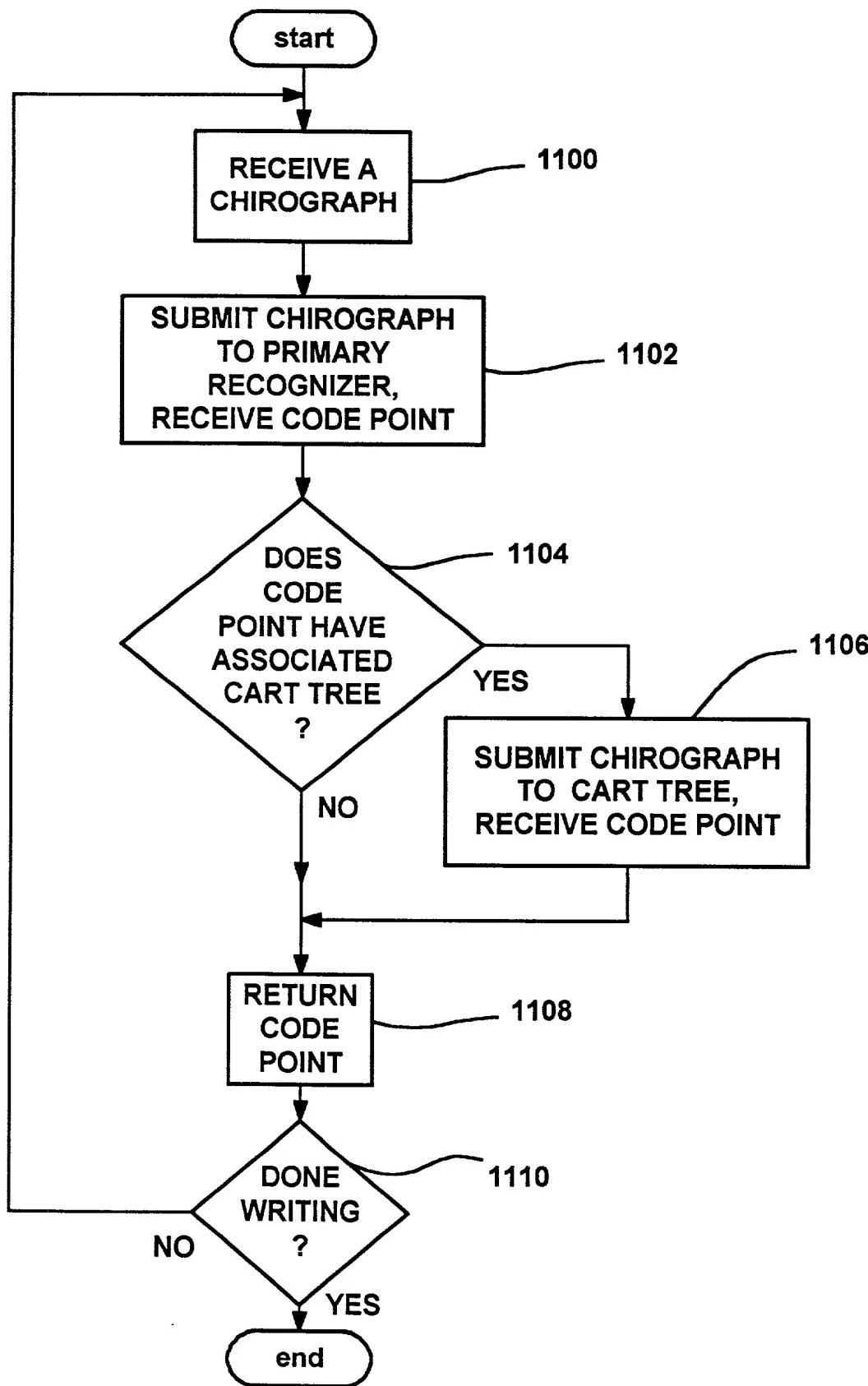


FIG. 10

FIG. 11



COMBINED DECLARATION AND POWER OF ATTORNEY

As a below named inventor, I hereby declare that:

My residence, post office address, and citizenship are as stated below next to my name.

I believe I am the original, first, and sole inventor (if only one name is listed below) or an original, first, and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled: "Method and Mechanism to Reduce Handwriting Recognizer Errors Using Multiple Decision Trees", the specification of which is attached hereto unless the following box is checked:

was filed on _____ as United States Application Serial No. _____ or PCT International Application No. _____ and was amended on _____ (if applicable).

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to patentability as defined in 37 C.F.R. § 1.56.

I hereby claim foreign priority benefits under 35 U.S.C. § 119(a)-(d) or § 365(b) of any foreign application(s) for patent or inventor's certificate, or § 365(a) of any PCT international application which designated at least one country other than the United States of America, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate, or PCT international application having a filing date before that of the application on which priority is claimed.

Prior Foreign Application(s) Priority Not Claimed

_____ (Number) _____ (Country) _____ (Day/Month/Year Filed)

_____ (Number) _____ (Country) _____ (Day/Month/Year Filed)

I hereby claim the benefit under 35 U.S.C. § 119(e) of any United States provisional application(s) listed below.

_____ (Application Number) _____ (Filing Date)

_____ (Application Number) _____ (Filing Date)

I hereby claim the benefit under 35 U.S.C. § 120 of any United States application(s), or § 365(c) of any PCT international application(s) designating the United States of America, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT International application(s) in the manner provided by the first paragraph of 35 U.S.C. § 112, I acknowledge the duty to disclose information which is material to patentability as defined in 37 C.F.R. § 1.56 which became available between the filing date of the prior application(s) and the national or PCT international filing date of this application.

_____ (Application Number) _____ (Filing Date) _____ (Status - patented, pending, abandoned)

_____ (Application Number) _____ (Filing Date) _____ (Status - patented, pending, abandoned)

As a named inventor, I hereby appoint the following attorney to prosecute this application and to transact all business in the Patent and Trademark Office connected therewith:

Albert S. Michalik, Reg. No. 37,395

Address all telephone calls to Albert S. Michalik at telephone number (425) 836-3030.

Address all correspondence to The Law Offices of Albert S. Michalik
704 - 228th Avenue NE
Suite # 193
Redmond, WA 98053

I hereby declare that all statements made herein of my own knowledge are true, that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full name of sole or first inventor: Gregory N. Hullender

Inventor's signature

Date June 6, 1997

Country of Citizenship: USA

Residence: 12211 NE 97th Street, Kirkland, Washington 98033

Post Office Address: Same

Full name of second joint inventor, if any: John R. Bennett

Inventor's signature

Date 6/16/97

Country of Citizenship: USA

Residence: 20724 NE 38th Street, Redmond, Washington 98053

Post Office Address: Same

Full name of third joint inventor, if any: Patrick M. Haluptzok

Inventor's signature

Date 6-6-97

Country of Citizenship: USA

Residence: 15244 NE 73rd Street, Redmond, Washington 98052

Post Office Address: Same

Additional inventors are being named on separately numbered sheets attached hereto.

1200 declaration